PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-236464
(43)Date of publication of application : 31.08.2001

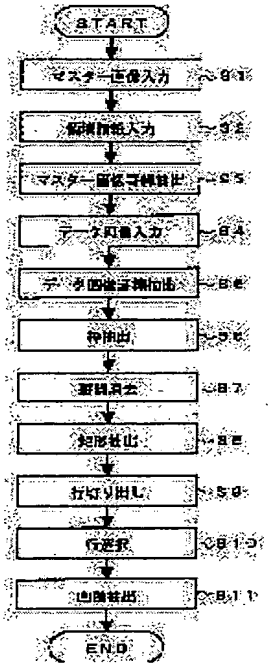
(51)Int.Cl. G06K 9/20
G06T 7/00

(21)Application number : 2000-048603
(22)Date of filing : 25.02.2000

(71)Applicant : RICOH CO LTD
(72)Inventor : BESSHO GORO

(54) METHOD AND DEVICE FOR CHARACTER EXTRACTION AND STORAGE MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To accurately extract even a character which sticks out from a frame on a document.
SOLUTION: The frame is recognized (S6), and the ruled lines constituting the frame are erased (S7). Rectangles of black pixel connecting components in an area obtained by expanding the frame are extracted from the document image having the ruled lines erased (S8) and are integrated to extract a character string area (line) (S9). Only the character area, which can be considered to belong to the frame, is selected from the extracted character string area (S10), and its image is cut out of the document image having the ruled lines erased (S11).



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998, 2000 Japan Patent Office

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－236464
（P2001－236464A）
(43)公開日　平成13年８月31日(2001.8.31)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'(参考)
Ｇ０６Ｋ　9/20	３４０	Ｇ０６Ｋ　9/20	３４０Ｌ	５Ｂ０２９
			３４０Ｋ	５Ｌ０９６
Ｇ０６Ｔ　7/00		Ｇ０６Ｆ　15/70	３３０Ｑ	

審査請求　未請求　請求項の数７　ＯＬ　（全７頁）

(21)出願番号　特願2000－48603(P2000－48603)

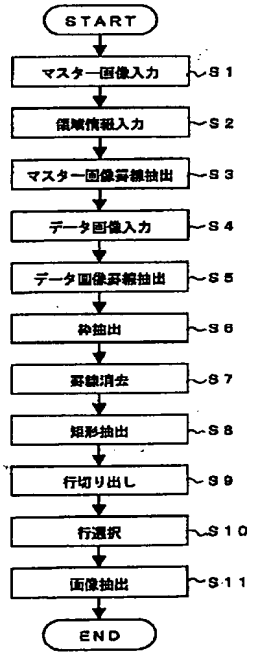
(22)出願日　平成12年２月25日(2000.2.25)

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　別所　吾朗
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(74)代理人　100073760
弁理士　鈴木　誠　（外１名）

Ｆターム(参考)　5B029 AA01 BB02 CC28 EE12
5L096 BA17 CA24 EA15 EA35 FA03
FA25 FA44 GA08 HA07

(54)【発明の名称】　文字抽出方法、文字抽出装置及び記憶媒体

(57)【要約】
【課題】　帳票上の枠からはみだした文字も正確に抽出する。
【解決手段】　枠を認識し（Ｓ６）、その枠を構成する罫線を消去する（Ｓ７）。罫線消去後の帳票画像より、枠を拡大した領域内の黒画素連結成分の矩形を抽出し（Ｓ８）、それら矩形の統合によって文字列領域（行）を抽出する（Ｓ９）。抽出された文字列領域の中から枠に属するとみなし得る文字列領域だけを選択し（Ｓ１０）、その画像を罫線消去後の帳票画像より切り出す（Ｓ１１）。



1

【特許請求の範囲】

【請求項１】　２値画像中の罫線によって囲まれた枠を認識する第１ステップと、この第１ステップで認識された枠を構成する罫線の黒画素を消去する第２ステップと、この第２ステップによって罫線の黒画素を消去された２値画像から、前記第１ステップで認識された枠を拡大した領域の内部の黒画素連結成分の矩形を抽出する第３ステップと、この第３ステップによって抽出された矩形を統合して文字列領域を抽出する第４ステップと、前記第１ステップによって認識された枠に関連して前記第４ステップによって抽出された文字列領域の中で当該枠に属するとみなし得る文字列領域を選択する第５ステップとを含むことを特徴とする文字抽出方法。

【請求項２】　前記第５ステップにおいて、文字列領域が枠に属するか否かを、当該文字列領域の当該枠の領域と重なる部分の面積割合に基づいて判定することを特徴とする請求項１記載の文字抽出方法。

【請求項３】　前記第３ステップにおいて、枠の拡大量を当該枠の短辺の長さを基準にして決定することを特徴とする請求項１記載の文字抽出方法。

【請求項４】　２値イメージデータを入力する第１手段と、この第１手段によって入力された２値イメージデータから罫線によって囲まれた枠を認識する第２手段と、この第２手段によって認識された枠を構成する罫線の黒画素を前記入力された２値イメージデータより消去する第３手段と、この第３手段によって罫線の黒画素が消去された後の２値イメージデータより、前記第２手段によって認識された枠を拡大した領域内の黒画素連結成分の矩形を抽出する第４手段と、この第４手段によって抽出された矩形を統合して文字列領域を抽出する第５手段と、前記第２手段によって認識された枠に関連して前記第５手段によって抽出された文字列領域の中で当該枠に属するとみなし得る文字列領域を選択する第６手段とを具備することを特徴とする文字抽出装置。

【請求項５】　前記第６手段において、文字列領域が枠に属するか否かを、当該文字列領域の当該枠の領域と重なる部分の面積割合に基づいて判定することを特徴とする請求項４記載の文字抽出装置。

【請求項６】　前記第４手段において、枠の拡大量を当該枠の短辺の長さを基準にして決定することを特徴とする請求項４記載の文字抽出装置。

【請求項７】　請求項１、２又は３記載の文字抽出方法の各ステップの処理をコンピュータに実行させるためのプログラムが記録されたことを特徴とするコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、画像処理の分野に係り、特に、文字認識装置などにおいて、表や帳票などの罫線に囲まれた枠の内部に記入された文字を抽出する

2

技術に関する。

【０００２】

【従来の技術】帳票などの罫線に囲まれた枠に記入された文字を抽出する方法として、特開平９－１６１００７号公報に開示されているように、罫線矩形を抽出し、罫線矩形の外側の座標を用いて枠を認識し、罫線画素を消去してから枠内の黒画素連結成分の矩形を抽出し、抽出した矩形の中で枠に接した矩形を除去し、残った矩形を用いて枠内の文字を切り出す方法が知られている。

【０００３】

【発明が解決しようとする課題】帳票において枠内に記入された文字が枠からはみ出していることがある。枠からはみ出した文字に対応した黒画素連結成分の矩形は、前記従来方法によれば枠に接触した矩形として除去されてしまう結果、枠からはみ出した文字を正常に抽出することができない。

【０００４】よって、本発明の目的は、枠からはみ出した文字も正常に抽出可能な文字抽出方法及び装置を提供することにある。

【０００５】

【課題を解決するための手段】上記目的を達成するため、本発明においては、２値画像より罫線によって囲まれた枠を認識し、認識した枠を構成する罫線の黒画素を消去し、罫線の黒画素を消去後の２値画像より、認識された枠を拡大した領域内の黒画素連結成分の矩形を抽出し、その矩形を統合して文字列領域を抽出し、認識された枠に関連して抽出された文字列領域の中で当該枠に属するとみなし得る文字列領域を選択する。また、文字列領域が枠に属するか否かの判定を、当該文字列領域の当該枠の領域と重なる部分の面積割合に基づいて行う。また、枠を拡大した領域の大きさを、枠の短辺の長さに応じて決定する。

【０００６】このような本発明の特徴及びその他の特徴について、実施の形態に関連して以下に詳述する。

【０００７】

【発明の実施の形態】以下、添付図面を参照し、本発明の実施の形態について説明する。図１は、本発明の実施の一形態である文字抽出装置のブロック構成の一例を示すブロック図である。図２は、この文字抽出装置における処理の流れを示すフローチャートである。また、また、図４乃至図７は処理の説明のため図である。

【０００８】この文字抽出装置は、例えば光学的文字認識装置の前処理部として用いられるもので、図１に見られるように、帳票の２値イメージデータを入力するためのスキャナなどの２値画像入力部１００と、これによって入力された２値イメージデータを蓄積するためのマスターイメージメモリ１０２及びデータイメージメモリ１０４と、罫線で囲まれた枠の認識に関連した領域情報入力部１０６、領域情報ファイル１０８、罫線抽出部１１０、アフィン変換係数メモリ１１２、罫線ファイル１１

４、罫線メモリ１１６、枠抽出部１１８及び枠領域メモリ１２０と、認識された枠を構成する罫線の黒画素を消去する（白画素に置き換える）ための罫線消去部１２２と、枠の拡大領域内の黒画素連結成分の矩形を抽出するための矩形抽出部１２４と、抽出された矩形の情報を記憶するための矩形メモリ１２６と、矩形を統合して文字列領域（行）を抽出する行切り出し部１２８と、抽出された文字列領域の情報を記憶するための行メモリ１３０と、抽出された文字列領域の中で枠に属するとみなされる文字列領域を選択し、その２値イメージデータを切り出す画像抽出部１３２と、切り出された２値イメージデータを記憶するための文字領域画像メモリ１３４とから構成される。

【０００９】以下、この文字抽出装置の動作について、処理の流れに沿って説明する。この文字抽出装置においては、帳票上の枠の認識のために、帳票のマスター画像（データの記入されていない空の帳票の画像）を利用するので、最初に２値画像入力部１００によって帳票のマスター画像の２値イメージデータを入力し、それをマスターイメージメモリ１０２に格納する（図２のステップＳ１）。

【００１０】そして、領域情報入力部１０６によって、マスター画像中の認識対象となる枠の位置や文字種などの情報を入力する（ステップＳ２）。具体的には、例えば、マスターイメージメモリ１０２内のマスター画像の２値イメージデータをディスプレイなどに表示し、マウスなどのポインティングデバイスを用いて認識対象の枠の位置などを指定する。入力された情報は領域情報ファイル１０８に格納される。なお、枠認識に利用されるのは、枠の位置に関する情報のみであるので、それ以外の情報の入力は省略可能である。

【００１１】次に、罫線抽出部１１０において、領域情報ファイル１０８内の枠の位置の情報を参照し、マスターイメージメモリ１０２内の２値イメージデータより認識対象の枠を構成する罫線を抽出し、その情報（始点、終点の座標など）を罫線ファイル１１４に格納する（ステップＳ３）。この罫線抽出は、例えば、主走査方向及び副走査方向の所定値以上の長さの黒ランを抽出し、所定の距離の範囲内にある黒ランを矩形に統合することによって各方向の緯線矩形を抽出する一般的な方法によって行うことができる。

【００１２】なお、ステップＳ１～Ｓ３は、新たなフォーマットの帳票を処理する場合にのみ行えばよいものである。例えば、同じフォーマットの帳票を多数枚連続して処理する時には、その１枚目の処理に先立ってステップＳ１～Ｓ３を行えばよい。この場合でも、そのフォーマットの帳票に関する領域情報及び罫線情報が領域情報ファイル１０８及び罫線ファイル１１４にそれぞれ登録されているならば、その情報を利用できるため、ステップＳ１～Ｓ３を改めて実行する必要はない。

【００１３】以上のようなマスター画像に関する領域情報及び罫線情報の登録が終了したならば（あるいは、既に得られている場合には、ステップＳ１～Ｓ３を行うことなく直ちに）、２値画像入力部１００によって、帳票のデータ画像（データが記入された帳票の画像）の２値イメージデータが入力され、データイメージメモリ１０４に格納される（ステップＳ４）。

【００１４】罫線抽出部１１０において、この入力された帳票のデータ画像上に罫線ファイル１１４に格納されている罫線の情報を投影することにより、データ画像上の認識対象の枠を構成する罫線を抽出し、その罫線の情報（始点、終点の座標など）を罫線メモリ１１６に格納する（ステップＳ５）。この際、マスター画像とデータ画像の間の位置ずれを計測し、その位置ずれの修正を行う。この位置ずれの計測には、例えば、両画像をシフトしながら、例えば罫線が十字に交差する点などの特徴点について画像間でパターンマッチングを行い、最も良好に一致する相対位置を求めるような方法を利用できる。また、位置ずれの修正のためのアフィン変換係数が求められてアフィン変化係数メモリ１１２に記憶され、このアフィン変換係数が各罫線の位置ずれの修正に適用される。データ画像の１枚１枚で位置ずれ量は異なるため、入力された各データ画像毎に、改めて位置ずれの計測とアフィン変換係数の算出が行われる。なお、位置ずれを問題にする必要がない場合には、位置ずれの計測と修正の処理を省略してよい。

【００１５】次に、枠抽出部１１８において、罫線メモリ１１６内の罫線情報を参照することにより、罫線によって囲まれた枠の領域を抽出し、その情報（始点と終点の座標など）を枠領域メモリ１２０に格納する（ステップＳ６）。

【００１６】次に、罫線消去部１２２において、罫線メモリ１１６内の罫線情報を参照し、抽出された罫線の黒画素を白画素に置換する処理を、データイメージメモリ１０４内のデータ画像の２値イメージデータに対して施す（ステップＳ７）。

【００１７】例えば、図４に示すような帳票のデータ画像上の「部番」の右側の枠が認識領域として指定された場合、図５に示すような枠１５０の領域が抽出され、また、この枠１５０を構成する罫線は消去される（図４との対応関係を分かりやすくするため、図５においては罫線の一部のみが消去されている）。

【００１８】次に矩形抽出部１２４は、黒画素連結成分の矩形を抽出するが、認識された枠の内部について矩形抽出を行うのではなく、枠を拡大した領域の内部について矩形抽出を行い、その情報（始点、終点の座標など）を矩形メモリ１２６に格納する（ステップＳ８）。例えば、図５に示す枠１５０の場合、図６に示すように枠１５０を拡大した領域１５３の内部について矩形抽出を行う。枠の拡大量は枠の短辺のサイズ(枠１５０のheigih

t）を基準にして決定される。このようにすれば、枠もしくは文字の大きさの違いに適応できる。ここに示す例は、横書きの帳票であり縦方向に文字のはみ出しが生じやすいため、枠１５０の高さheightの半分だけ、枠１５０を上下に拡張している。

【００１９】次に、行切り出し部１２８において、矩形メモリ１２６内の矩形情報を参照し、所定の距離範囲内にある矩形を統合することにより文字列領域（行）を抽出し、その情報（始点、終点の座標など）を行メモリ１３０に格納する（ステップＳ９）。日本語が用いられる帳票では、行の方向は縦方向と横方向の両方が考えられるが、多くの場合、帳票中の行はほとんどが横方向であるので、横方向に連続した矩形の統合により横方向の文字列領域を抽出すればよい。図６に示す枠の拡大領域１５３においては、図７に示す３つの文字列領域１６１，１６２，１６３が抽出される。

【００２０】次に、画像抽出部１３２において、枠領域メモリ１２０内の枠領域情報及び行メモリ１３０内の文字列領域情報を参照し、各枠に関して抽出された文字列領域の中で、その枠に属するとみなし得る文字列領域を選択し（ステップＳ１０）、選択した文字列領域の２値イメージデータをデータイメージメモリ１０４内の罫線消去後の２値イメージデータより切り出し、それを文字領域画像メモリ１３４に格納する（ステップＳ１１）。

【００２１】文字列領域が枠に属するとみなし得るか否かの判定は、例えば次のようにして行われる。

（１）拡大前の枠の内部に全体が包含される文字列領域は、枠に属する文字列領域と判定する。

（２）全体が拡大前の枠の外にある文字列領域は、枠に属する文字列領域ではないと判定する。

（３）上記条件のいずれにも該当しない文字列領域については、文字列領域の拡大前の枠の領域と重なる部分の、文字列領域全体対する面積割合を計算する。この面積割合が１／２以上（全体の半分以上が拡大前の枠と重なる）文字列領域は枠に属する文字列領域と判定し、面積割合が１／２未満の文字列領域は枠に属する文字列領域ではないと判定する。

【００２２】図７の文字列領域１６１，１６２，１６３のうちで、文字列領域１６１は上記条件（１）により枠１５０に属すると判定される。文字列領域１６３は、上記条件（２）により枠１５０に属さないと判定される。文字列領域１６２は、面積割合が１／２未満であるため、上記条件（３）により枠１５０に属さないと判定される。したがって、枠１５０の文字列領域としては文字列領域１６１だけか選択され、その画像が正しく切り出される。また、このような文字列（行）を単位として、枠に属するか否かの判定を行うことにより、分離文字の一部が誤って抽出されたり、抽出されなかったりする不都合を回避できる。枠１５０に属さないと判定される文字列領域１６２は、１つ下の「品名」の枠の処理時に、

その枠に属する文字列領域として選択され、その２値イメージデータが切り出されることは以上の説明から明らかである。

【００２３】このように、文字列領域１６２内の文字のように、枠からはみ出した文字も正しく抽出することができる。なお、枠に属する文字列領域が２つ以上抽出される場合には、それら文字列領域の画像を１つの文字列領域の画像として統合して抽出するようにしてもよい。

【００２４】本発明の他の実施の形態によれば、マスター画像に関して予め登録した情報を利用せず、帳票のデータ画像から直接的に枠が認識される。図１を援用して説明すれば、罫線抽出部１１０において、例えば、データイメージメモリ１０４内の２値イメージデータをスキャンし、主走査方向及び副走査方向の所定値以上の長さの黒ランを抽出し、抽出した各方向の黒ランに対し、所定の距離範囲内にある黒ランを統合する処理を行うことによって罫線の矩形を抽出し、その始点と終点の座標を罫線メモリ１１６に格納する。以下、前述の実施の形態の場合と同様の処理を行って枠の文字の抽出を行う。なお、領域情報入力部１０６と同様に、帳票のデータ画像をディスプレイに表示し、マウスなどを用いて認識対象の枠の位置を指定し、指定された枠を構成する罫線のみを抽出するようにしてもよい。このような実施の形態は、予め形式の定まっていない不定形の帳票を処理する場合に有効である。

【００２５】前述のような本発明による文字抽出装置及びその処理は、必ずしも専用のハードウェアによって実現される必要はなく、例えば図３に示すようなＣＰＵ２０１、メモリ２０２、ハードディスク２０３、入力装置（イメージスキャナ、キーボード、マウスなど）２０４、ディスプレイ２０６、各種記憶媒体（磁気ディスク、光ディスク、光磁気ディスク、メモリカードなど）２０７の読み書きのための媒体ドライブ２０５、外部の機器やネットワークとの通信のための通信装置（モデム、ネットワークアダプタなど）２０８などをバス２１０で接続した一般的な構成のコンピュータを利用して、ソフトウェアにより実現することもできる。この場合、文字抽出装置の各部の機能をコンピュータ上で実現するためのプログラム、換言すれば、文字抽出のための各処理ステップをコンピュータ上で実行するためのプログラムが、例えば、それが記録された記憶媒体２０７から媒体ドライブ２０５によってメモリ２０２に読み込まれたり、外部装置より通信装置２０８によってメモリ２０２に読み込まれたり、あるいはハードディスク２０３からメモリ２０２に読み込まれ、ＣＰＵ２０１により実行される。このプログラムを、それを固定記憶させた半導体ＲＯＭとして実装してもよい。このようなプログラムが記録された各種記憶媒体２０７や半導体ＲＯＭなども本発明に包含されるものである。また、帳票の２値イメージデータは、例えば入力装置２０４に含まれるスキャナ

7

によって読み込まれたり、記憶媒体２０７から読み込まれたり、あるいは通信装置２０８によって外部機器から入力される。

【００２６】

【発明の効果】以上の詳細な説明から明らかなように、本発明の文字抽出方法又は装置によれば、枠から文字がはみ出して記入されている場合であっても、正確な文字抽出が可能となる。したがって、本発明の文字抽出方法又は装置を文字認識の前処理に適用すれば、枠からはみ出して記入された文字も正確な文字認識が可能となる。また、本発明の記憶媒体を用いれば、一般的なコンピュータを利用して上に述べたような正確な文字抽出が可能となる、等々の効果を得られる。

【図面の簡単な説明】

【図１】本発明による文字抽出装置のブロック構成の一例を示すブロック図である。

【図２】図１の文字抽出装置における処理の流れを示すフローチャートである。

【図３】本発明をソフトウェアにより実施するために利用されるコンピュータの一例を示すブロック図である。

【図４】帳票のデータ画像と、その枠の一例を示す図である。

【図５】罫線消去後のデータ画像上の枠の領域を示す図である。

【図６】枠を拡大した領域を示す図である。

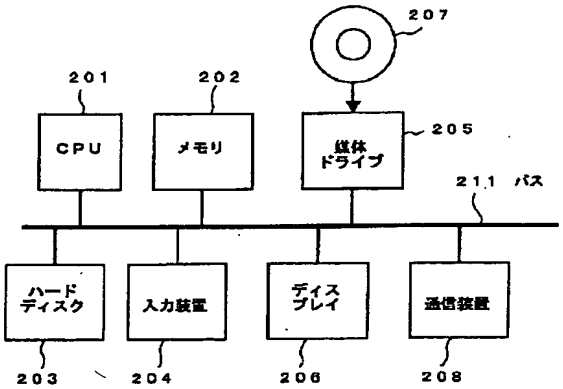
8

【図７】枠を拡大した領域から抽出される文字列領域を示す図である。

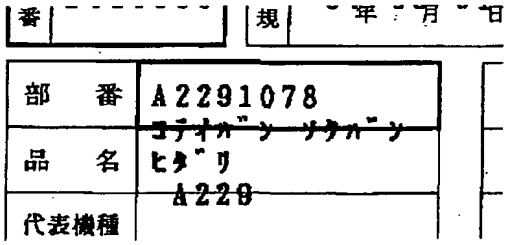
【符号の説明】

１００　２値画像入力部
１０２　マスターイメージメモリ
１０４　データイメージメモリ
１０６　認識領域情報入力部
１０８　領域情報ファイル
１１０　罫線抽出部
１１２　アフィン変換係数メモリ
１１４　罫線ファイル
１１６　罫線メモリ
１１８　枠抽出部
１２０　枠領域メモリ
１２２　罫線消去部
１２４　矩形抽出部
１２６　矩形メモリ
１２８　行切り出し部
１３０　行メモリ
１３２　画像抽出部
１３４　文字領域メモリ
１５０　枠
１５３　枠の拡大領域
１６１，１６２，１６３　文字列領域

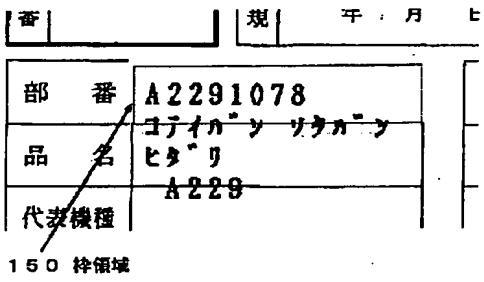
【図３】



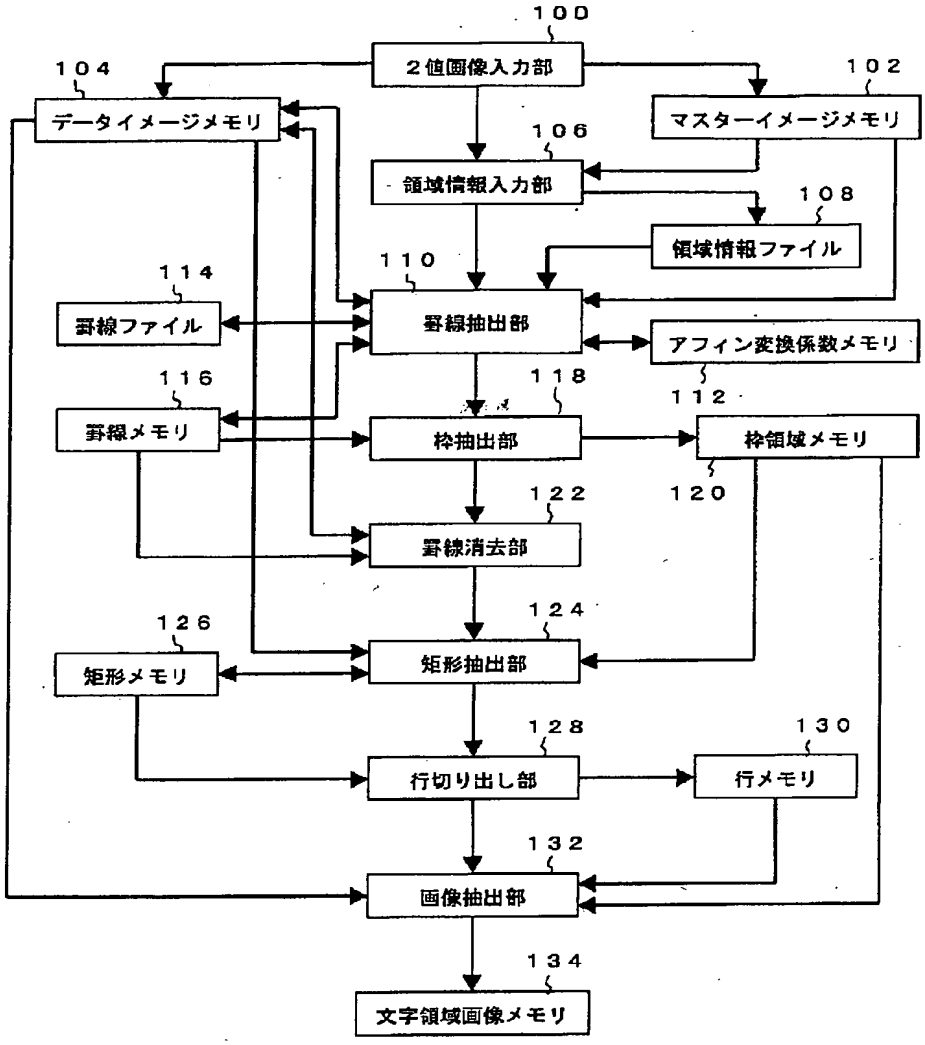
【図４】



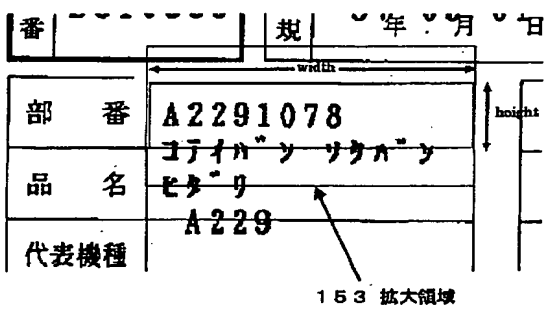
【図５】



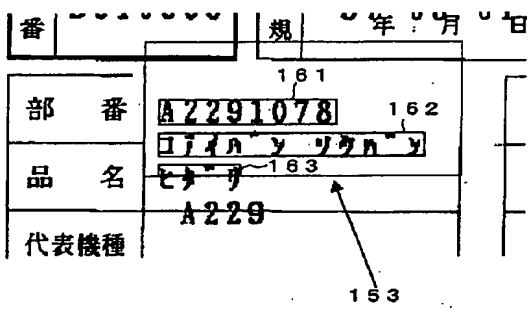
【図１】


100
２値画像入力部
104
データイメージメモリ
102
マスターイメージメモリ
106
領域情報入力部
108
領域情報ファイル
110
罫線抽出部
114
罫線ファイル
アフィン変換係数メモリ
112
116
罫線メモリ
118
枠抽出部
枠領域メモリ
120
122
罫線消去部
124
126
矩形メモリ
矩形抽出部
128
130
行切り出し部
行メモリ
132
画像抽出部
134
文字領域画像メモリ

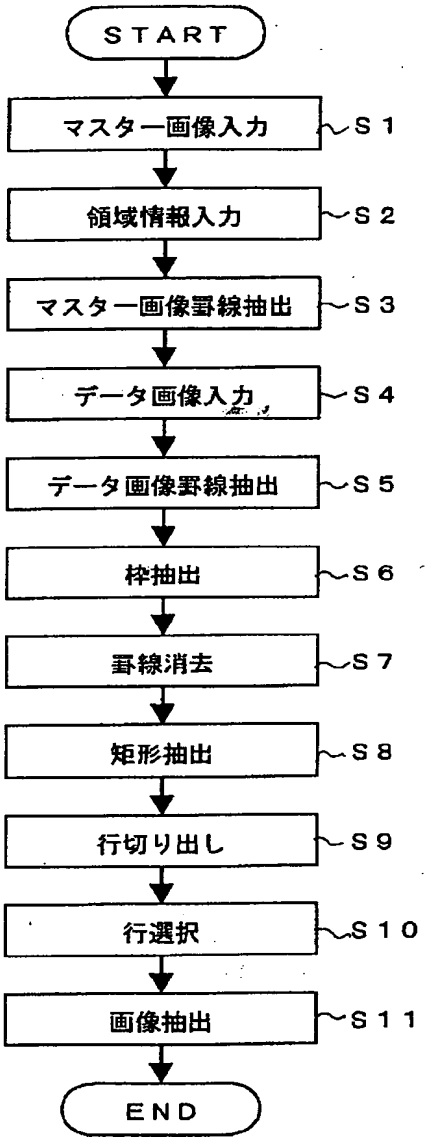
【図６】


width
height
部　番
A2291078
品　名
A229
代表機種
153 拡大領域

【図７】


161
部　番
A2291078
162
品　名
163
A229
代表機種
153

【図２】


START
マスター画像入力
S１
領域情報入力
S２
マスター画像罫線抽出
S３
データ画像入力
S４
データ画像罫線抽出
S５
枠抽出
S６
罫線消去
S７
矩形抽出
S８
行切り出し
S９
行選択
S１０
画像抽出
S１１
END